D172–D177 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 31 October 2013*
doi:10.1093/nar/gkt1036

BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human blood cells

Diego Chacon, Dominik Beck, Dilmi Perera, Jason W. H. Wong* and John E. Pimanda*

Lowy Cancer Research Centre and the Prince of Wales Clinical School, University of New South Wales, Sydney, NSW 2052, Australia

Received August 15, 2013; Revised September 23, 2013; Accepted October 9, 2013

ABSTRACT

The BloodChIP database (http://www.med.unsw.edu.au/CRCWeb.nsf/page/BloodChIP) supports exploration and visualization of combinatorial transcription factor (TF) binding at a particular locus in human CD34-positive and other normal and leukaemic cells or retrieval of target gene sets for user-defined combinations of TFs across one or more cell types. Increasing numbers of genome-wide TF binding profiles are being added to public repositories, and this trend is likely to continue. For the power of these data sets to be fully harnessed by experimental scientists, there is a need for these data to be placed in context and easily accessible for downstream applications. To this end, we have built a user-friendly database that has at its core the genome-wide binding profiles of seven key haematopoietic TFs in human stem/progenitor cells. These binding profiles are compared with binding profiles in normal differentiated and leukaemic cells. We have integrated these TF binding profiles with chromatin marks and expression data in normal and leukaemic cell fractions. All queries can be exported into external sites to construct TF–gene and protein–protein networks and to evaluate the association of genes with cellular processes and tissue expression.

INTRODUCTION

Transcription factors (TFs) and the *cis*-regulatory sequences to which they bind form the building blocks of gene regulatory networks that govern gene expression and give cells their unique identity (1). Stem cells have the capacity to regenerate themselves and also to give rise to progeny with increasingly specialized function and restricted proliferative capacity (2). Haematopoiesis is one of the best described developmental systems in vertebrates and serves as a model system for the multi-lineage differentiation of adult stem cells. However, little is known about the components and hierarchy of the gene regulatory network that controls the identity of human adult haematopoietic stem cells (HSCs). Understanding the ground state of the transcriptional network in HSCs is the first step to better understanding how these networks change as HSCs respond to external cues and proliferate or differentiate. Stem cell signatures are also corrupted and persist in leukaemic cells and confer an adverse prognosis (3,4). Knowledge of the components and hierarchy of the normal HSC transcriptional network may provide clues to dismantling stem cell networks in leukaemic cells.

TFs often act in 3-D space as components of multiprotein complexes that bind distal regulatory elements and interact with promoters to initiate, augment or suppress gene expression in conjunction with a plethora of chromatin modifiers and transcription initiation/elongation factors (5). In the post-genomics era, classical methods to screen for potential gene-regulatory regions, such as DNaseI hypersensitivity mapping by Southern blotting of selected loci, have been superseded by new genome-wide techniques such as Chromatin Immunoprecipitation (ChIP) and DNaseI hypersensitivity assays coupled to high throughput sequencing (ChIP-seq and DNaseI-seq, respectively) (6,7). These large-scale screening techniques help rapidly identify the positions of a large number of candidate regulatory elements and genome-wide binding patterns for multiple TFs.

A core set of seven TFs—FLI1, ERG, GATA2, RUNX1, SCL/TAL1, LYL1 and LMO2—work in combination to regulate gene expression in haematopoietic stem/progenitor cells (HSPCs) (8). This heptad of TFs also contributes to stem cell-like signatures in leukaemic cells, and their presence impacts adversely on patient

*To whom correspondence should be addressed. Tel: +61 293 851 003; Fax: +61 293 851 510; Email: jpimanda@unsw.edu.au
Correspondence may also be addressed to Jason W. H. Wong. Tel: +61 293 858 796; Fax: +61 293 851 510; Email: jason.wong@unsw.edu.au

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

survival (4). Knowledge of how these TFs cooperate with each other and interact with other lineage-specific TFs to regulate gene expression during HSPC differentiation is vital to improving our understanding of normal blood development. To this end, we have recently generated high-quality ChIP-seq data for these seven haematopoietic TFs in primary human $CD34^+$ blood stem and progenitor cells (9). To optimize the utility of these data, we have integrated our ChIP-seq and RNA-seq data sets with chromatin accessibility marks from the Human Epigenome Atlas (6) and Encyclopaedia of DNA Elements (ENCODE) (10) as well as published ChIP-seq and expression data from normal (11–13) and leukaemic stem cell fractions (14). All analysed and filtered data sets can be exported for downstream analysis. To facilitate this process, we have also created links to external resources such as the UCSC genome browser to visualize data, Cytoscape (15) and STRING (16) to construct TF–gene and protein–protein networks and the Gene Expression Atlas (17) to query expression in other data sets and Gene Set Enrichment Analysis (GSEA) (18) to associate a list of genes with cellular processes.

MATERIALS AND METHODS

Data sources

All data sources are summarized in Supplementary Table S1 and a schematic diagram illustrating data sets and features of BloodChIP is shown in Figure 1.

TF ChIP-seq

Genome-wide TF binding profiles for Human CD34+ HSPC (GSE45144), megakaryocytes (GSE24674), the acute myeloid leukemia (AML) cell line SKNO-1 (GSE23730) and K562 (GSE24779, GSE29196, GSE31477) were obtained from the Gene Expression Omnibus (GEO) (9,12,19–21). To standardize the way TF binding sites are obtained from the data set, peak calling was performed using an improved automated pipeline that we used previously to analyse the Human CD34+ HSPC data set (9). Briefly, sequencing reads were aligned to the hg19 reference genome using Burrows-Wheeler Aligner (BWA) (22). Peak calling was performed using three peak calling algorithms, HOMER (23), MACS (24) and SPP (25). Only peaks that were called by two or more of the algorithms are kept as part of the final peak list. Aligned reads were also extended to 200 bp to generated TF-binding genome coverage profiles for visualization on the UCSC genome browser. A schematic diagram illustrating the pipeline is shown in Figure 2.

To obtain a uniform set of TF binding sites across all TFs and across all cell lines, all peaks were merged. This was achieved by sequentially merging overlapping peaks across pairs of data set until all data sets have been merged. Each region was then marked as either bound or unbound for each TF data set. TF binding peaks were mapped to nearby genes using the genomic regions enrichment of annotations tool (GREAT) (26), which defines genomic neighbourhoods for TF-bound peaks by assigning weights to flanking genes based on their distance to the peak.

Histone ChIP-seq

Histone ChIP-seq profiles from mobilized CD34+ cells and K562 cells were obtained from the NIH Roadmap Epigenomics Project (GSE18927) (6) and the ENCODE project (GSE29611) (21), respectively. Histone profiles for H3K27Ac, H3K4me1 and H3K4me3 were analysed, and binding was quantified by counting the number to tags mapping to 1.5 kbs adjacent to the centre of each TF binding site. A read count of greater than 10 tags was used as evidence for the presence of a particular histone mark, which has been shown previously to adequately distinguish true signal from noise (9).

Gene expression

Genome-wide microarray expression data for human normal CD34+ cells (GSE30029) (27), SKNO-1 cells (GSE34594) (28) and K562 (GSE28135) (29) were obtained from GEO (30), and expression data for megakaryocytes (E-TABM-633) (11) was obtained from ArrayExpress (31). Illumina expression arrays were used in all of the above studies, the non-normalized expression data from each data set was log2 transformed and quantile normalized across all samples together for all cell types. Where multiple microarray probes are available for one gene, the probe with the highest average expression value across all samples was selected to represent the expression of the gene.

Genome-wide expression data comparing the stem and progenitor sub-fractions of human CD34+ cells of normal donors and AML patients was obtained from GEO (GSE24006) (14). Expression was measured on an Affymetrix array. Robust Multi-chip Average approach was used for normalization and expression levels summarized. Probes were mapped to respective genes and expression values for each gene were stored in the database.

DNaseI hypersensitive site and digital genomic footprints

For analysis of chromatin accessibility and digital genomic footprints, annotations for mobilized CD34+ cells and K562 cells (GSM646567) were obtained from previously published data (32). A TF bound region was annotated as chromatin accessible if the peak region overlapped with a DNase I hypersensitive site (HS) region. Furthermore, the number of digital genomic footprints was counted within each DNase I HS region.

Database implementation, web interface and availability

All processed data were stored in a local MySQL (version 5.5.8) relational database. A web interface has been designed to provide user-friendly access for database query and visualization of TF binding and gene expression through PHP-based scripts. The web interface is accessible at: http://www.med.unsw.edu.au/CRCWeb.nsf/page/bloodChIP. BloodChIP also supports REST

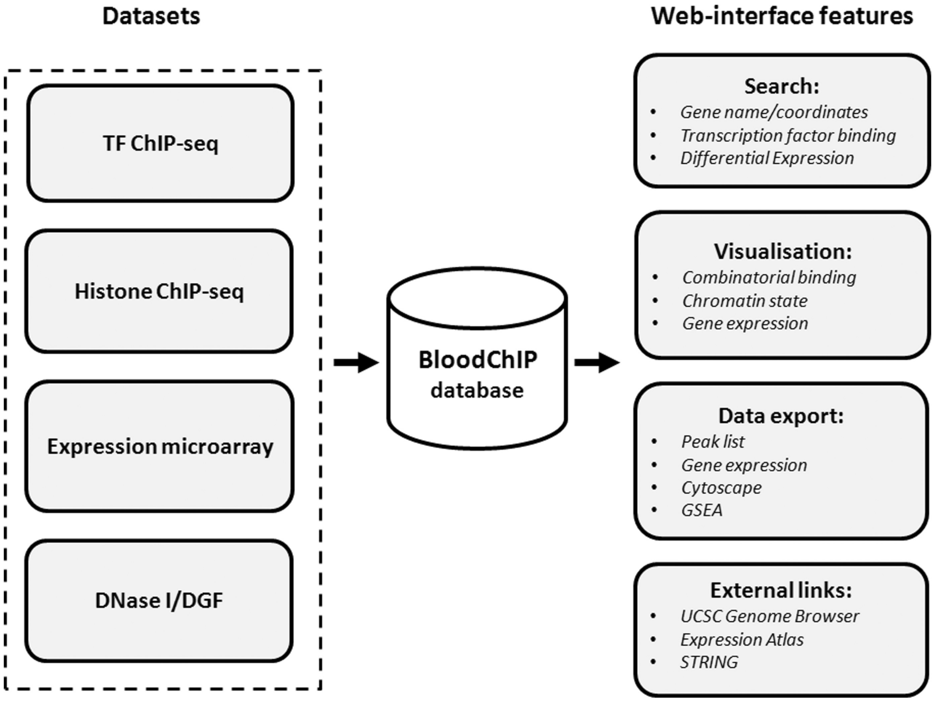

Figure 1. Schematic summarizing data currently held by the BloodChIP database and features of the web interface. The database integrates TF ChIP-seq, histone ChIP-seq, DNase I/digital genomic footprinting (DGF) and expression microarray data. The web interface provides methods to query and visualize data and further links to external databases for further data analysis.

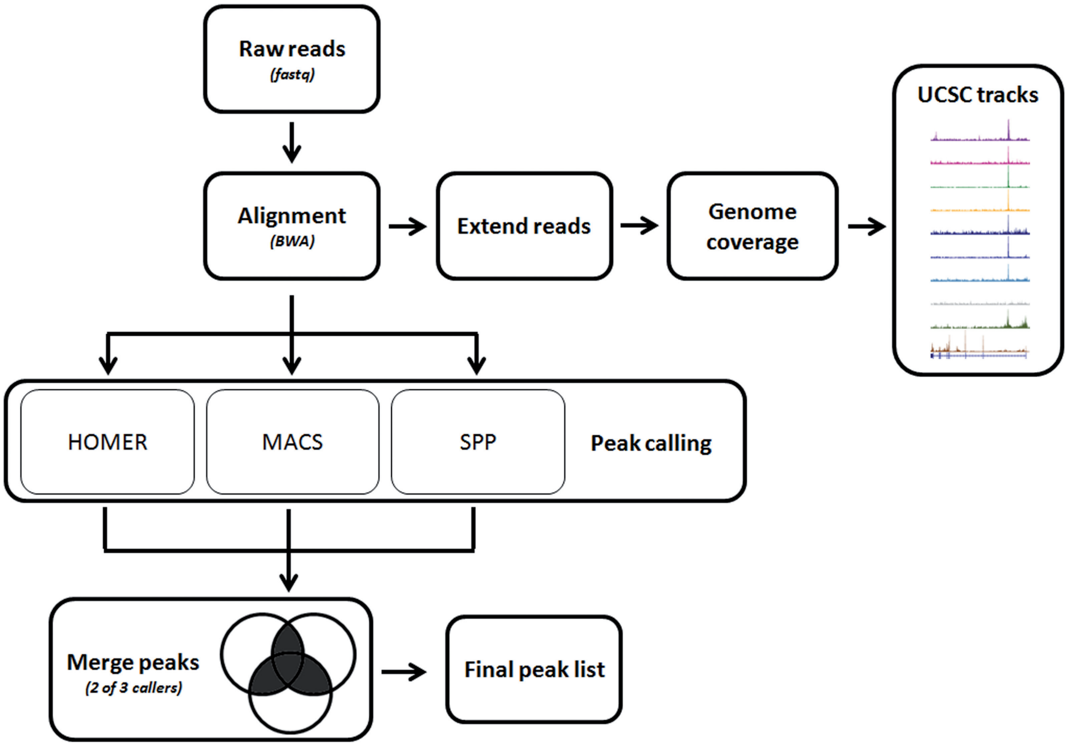

Figure 2. Schematic illustrating the ChIP-seq data analysis pipeline used for processing of ChIP-seq data sets used to populate the BloodChIP database. The two outputs from the pipeline are the genome coverage profiles for visualization in UCSC and a list of transcription factor binding sites.

(REpresentational State Transfer) Web Services to enable programmatic access to the database. A tutorial is provided at the BloodChIP web page describing the use of the REST service. All data from BloodChIP are also available for download as SQL files.

RESULTS: DATABASE FEATURES

General web interface

By default, the web interface to the database displays a main table that lists all 23 804 genes in the human genome as a sortable table (Supplementary Figure S1). For each gene, the total number of heptad ChIP-seq peaks associated with the gene is shown. A link to the UCSC genome browser is also provided to allow visualization of tracks for TF ChIP-seq, histone profiles and DNaseI HS at the respective gene locus. Clicking on the row of each gene will display the distribution of expression for a particular gene in CD34+, megakaryocytes, SKNO-1 and K562 cells in the form of a box plot on the right of the main table. This will also display checkboxes that indicate TF binding associated with the particular gene.

To explore TF ChIP-seq peaks associated with a particular gene, the row containing the gene can be expanded. The genomic coordinates of each peak is shown, along with red boxes to indicate the binding of a particular TF from a particular cell line at the particular peak. A blue box indicates the absence of binding and a white box indicates that the TF-binding profile is not present for the particular cell type. The presence of each of the histone marks and DNase I hypersensitivity is also shown. Finally, the number of digital genomic footprints is also indicated. The binding status of a maximum of seven TFs in three cell types can be displayed on the web interface at any one time. The specific TFs and cell types displayed can be customized by the user using the display filter above the main table.

Querying the database

Two methods are available for querying the database using the web interface, namely using gene symbol/genomic coordinates and TF binding. For queries using gene/genomic coordinate, a search dialogue is provided to accept searches using the gene symbol, Refseq ID or genomic coordinates in bed format. A user can either paste in the query list or upload this as a file.

For retrieval of genes regulated by particular TF, a user can select specific TFs for the different cell types available in the database. A query can be constructed with any combination of TFs to identify either genes that are combinatorially bound by multiple TFs ("AND") or bound by any of the TFs ("OR"). Furthermore, an option is available to filter the retrieved genes based on differential gene expression between any of the cell types. A cut-off can be specified to retain only genes that are above or below a certain fold change threshold.

Links to external database and tools

To facilitate downstream analysis and data visualization, hyperlinks taking advantage of external APIs are provided. To further investigate the expression of a single gene in other experiments, a link is provided to the Gene Expression Atlas, displaying condition and experiment-specific expression patterns of the gene of interest. The Gene Expression Atlas is useful for the identification of publically available gene expression data sets that can be used for further analysis and validation of hypotheses.

To explore protein–protein interactions of the set of genes displayed in main table, a link is provided to the STRING database (16). The STRING database can be used to reveal known and predicted physical and functional protein interactions. This is particularly useful for the discovery of major protein–protein interaction hubs and for the prioritization of genes for further analysis.

Exporting data

The web interface provides four methods for the exporting of data from a queried set of genes. Firstly, TF binding information can be exported for individual genes in tab-delimited format. Secondly, the normalized expression values of all genes retrieved by a query for all samples can be exported. This facilitates external analysis such as clustering of genes or the generation of heatmaps. Thirdly, the set of genes retrieved from the query can be exported as a gene set for use in GSEA. Finally, to allow the visualization of the TF–gene regulatory network, a file can be exported in a format that can be immediately visualized and analysed using the Cytoscape software package.

Utility of BloodChIP: a working example

If the user is interested in retrieving TF binding data at the *RUNX1* locus, a query for *RUNX1* is initiated by typing in the gene name or gene coordinates. The default settings retrieve all combinations of binding sites with one or more TF peaks that have been mapped to a locus by GREAT (26) (Supplementary Figure S2A). The resulting view shows peak coordinates in hCD34, Megakaryocytes and AML cells with a link to the UCSC browser and a checkerboard view of TF(s) bound to this region (Supplementary Figure S2B). The Chr21: 36398905-36399463 interval, which is bound by all seven TFs and has active chromatin marks, corresponds to the *Runx1*+23 stem cell enhancer in mice (33).

This view also permits easy visualization of comparative binding profiles at these or other regions in primary megakaryocytes (12) and AML cells (20). The gene expression view (Supplementary Figure S2C) to the right shows RUNX1 expression across HSCs, multi-potent progenitors (MPP), common myeloid progenitors (CMP), granulocyte–monocyte progenitors (GMP) or megakaryocyte–erythroid progenitor (MEP) fractions as well as in AML leukaemic stem cells (LSC; Lin-/CD34+/38-/CD90-), AML leukaemic progenitor cells (Lin-/34+/38+) and AML blasts (Lin-/34-) (14), megakaryocytes and AML cells. Had the biological function of the +23 enhancer not been known, this region would have been the prime candidate for functional testing as a regulator of a gene that is both important for normal blood development and is mutated in leukaemia.

A tab at the top left corner permits easy export of data contained in this view.

Alternatively, if the user wished to retrieve all targets for RUNX1 alone or in combination with one or more TFs, the appropriate options corresponding to the particular cell type(s) of interest can be selected to yield a list of genes that can either be viewed on UCSC or exported to retrieve coordinates. For example, if the selects RUNX1 (CD34) and FLI1 (CD34), the user will retrieve sites with combinatorial binding for RUNX1 and FLI1 in CD34+ cells. If on the other hand RUNX1 (CD34) or FLI1 (CD34) is chosen, the user will retrieve all RUNX1 coordinates and FLI1 coordinates in CD34+ cells. Protein–protein and TF–gene interactions for this list can also be visualized by following the adjacent tabs to STRING (Supplementary Figure S2E) and Cytoscape (Supplementary Figure S3A). Data can also be exported into GSEA (Supplementary Figure S3B) to evaluate associations with cellular processes or for other applications such as generation of heatmaps using a tool of choice (Supplementary Figure S3C). Another feature of the database is the function to filter outputs based on differential expression between normal HSCs and more differentiated normal blood subsets or normal HSCs and leukaemic stem cell fractions. Binding profiles and binding coordinates of each gene on the list can be accessed and compared between normal HSCs and leukaemic cell lines.

DISCUSSION

Combinatorial interactions of TFs are key determinants of cell identity (34). We have recently generated genome-wide high resolution binding profiles for seven key haematopoietic TFs in primary human CD34+ haematopoietic stem progenitor cells (HSPCs) (9). We have now integrated combinatorial TF binding data with quantitative gene expression, histone modification and digital genomic footprinting data in these cells from the Human Epigenome Atlas (6) and ENCODE (10) and created a user-friendly database that allows users to (i) Interrogate overlapping TF binding at a locus of interest, (ii) Establish binding coordinates and associated gene targets for TFs of interest, (iii) Compare binding profiles between cell types and (iv) Filter lists based on differentially expressed genes. Additional features include tools to construct visual maps of TF–gene and protein–protein interaction networks using links to Cytoscape (35) and STRING (16), respectively. The overall aim of this database is to facilitate enquiry into dynamic changes to the transcriptional network of HSPCs as cells differentiate into specific lineages or transform into leukaemic cells.

At present, we have included data sets for normal hCD34s, primary megakaryocytes and an AML cell line for which high-quality combinatorial TF binding data are available. As more TF binding data are published and where the quality of these data sets accord with ENCODE guidelines (36), BloodChIP will be expanded to accommodate more TF binding profiles in human CD34+ HSPCs as well as those in other normal human blood lineages and human leukaemic cell lines. We believe that BloodChIP, which integrates combinatorial TF binding with gene expression in normal and malignant cells, will be of particular interest to biologists and bioinformaticians working in the fields of haematopoiesis and leukaemia and also more generally to those working on gene regulation and stem cell biology.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Health and Medical Research Council of Australia; Australian Research Council; Leukaemia Foundation; Cancer Institute for New South Wales. Funding for open access charge: National Health and Medical Research Council of Australia.

Conflict of interest statement. None declared.

REFERENCES

1. Pimanda,J.E. and Gottgens,B. (2010) Gene regulatory networks governing haematopoietic stem cell development and identity. *Int. J. Dev. Biol.*, **54**, 1201–1211.
2. Reya,T., Morrison,S.J., Clarke,M.F. and Weissman,I.L. (2001) Stem cells, cancer, and cancer stem cells. *Nature*, **414**, 105–111.
3. Eppert,K., Takenaka,K., Lechman,E.R., Waldron,L., Nilsson,B., van Galen,P., Metzeler,K.H., Poeppl,A., Ling,V., Beyene,J. *et al.* (2011) Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.*, **17**, 1086–1093.
4. Diffner,E., Beck,D., Gudgin,E., Thoms,J.A., Knezevic,K., Pridans,C., Foster,S., Goode,D., Lim,W.K., Boelen,L. *et al.* (2013) Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. *Blood*, **121**, 2289–2300.
5. Edelman,L.B. and Fraser,P. (2012) Transcription factories: genetic programming in three dimensions. *Curr. Opin. Genet. Dev.*, **22**, 110–114.
6. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
7. Bernstein,B.E., Meissner,A. and Lander,E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
8. Wilson,N.K., Foster,S.D., Wang,X., Knezevic,K., Schutte,J., Kaimakis,P., Chilarska,P.M., Kinston,S., Ouwehand,W.H., Dzierzak,E. *et al.* (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*, **7**, 532–544.
9. Beck,D., Thoms,J., Unnikrishnan,A., Knezevic,K., Perera,D., O'Brien,T., Gottgens,B., Wong,J.W. and Pimanda,J. (2013) Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected circuit of coding and non-coding genes. *Blood*, **122**, e12–e22.
10. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
11. Watkins,N.A., Gusnanto,A., de Bono,B., De,S., Miranda-Saavedra,D., Hardie,D.L., Angenent,W.G., Attwood,A.P., Ellis,P.D., Erber,W. *et al.* (2009) A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, **113**, e1–e9.
12. Tijssen,M.R., Cvejic,A., Joshi,A., Hannah,R.L., Ferreira,R., Forrai,A., Bellissimo,D.C., Oram,S.H., Smethurst,P.A., Wilson,N.K. *et al.* (2011) Genome-wide analysis of simultaneous

GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. *Dev. Cell*, **20**, 597–609.

13. Novershtern,N., Subramanian,A., Lawton,L.N., Mak,R.H., Haining,W.N., McConkey,M.E., Habib,N., Yosef,N., Chang,C.Y., Shay,T. *et al.* (2011) Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**, 296–309.
14. Gentles,A.J., Plevritis,S.K., Majeti,R. and Alizadeh,A.A. (2010) Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *JAMA*, **304**, 2706–2715.
15. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
16. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P., von Mering,C. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–815.
17. Kapushesky,M., Adamusiak,T., Burdett,T., Culhane,A., Farne,A., Filippov,A., Holloway,E., Klebanov,A., Kryvych,N., Kurbatova,N. *et al.* (2012) Gene expression atlas update–a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **40**, D1077–D1081.
18. Subramanian,A., Kuehn,H., Gould,J., Tamayo,P. and Mesirov,J.P. (2007) GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics*, **23**, 3251–3253.
19. Trompouki,E., Bowman,T.V., Lawton,L.N., Fan,Z.P., Wu,D.C., DiBiase,A., Martin,C.S., Cech,J.N., Sessa,A.K., Leblanc,J.L. *et al.* (2011) Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell*, **147**, 577–589.
20. Martens,J.H., Mandoli,A., Simmer,F., Wierenga,B.J., Saeed,S., Singh,A.A., Altucci,L., Vellenga,E. and Stunnenberg,H.G. (2012) ERG and FLI1 binding sites demarcate targets for aberrant epigenetic regulation by AML1-ETO in acute myeloid leukemia. *Blood*, **120**, 4038–4048.
21. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
22. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
23. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
24. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
25. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
26. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
27. de Jonge,H.J., Woolthuis,C.M., Vos,A.Z., Mulder,A., van den Berg,E., Kluin,P.M., van der Weide,K., de Bont,E.S., Huls,G., Vellenga,E. *et al.* (2011) Gene expression profiling in the leukemic stem cell-enriched CD34+ fraction identifies target genes that predict prognosis in normal karyotype AML. *Leukemia*, **25**, 1825–1833.
28. Ptasinska,A., Assi,S.A., Mannari,D., James,S.R., Williamson,D., Dunne,J., Hoogenkamp,M., Wu,M., Care,M., McNeill,H. *et al.* (2012) Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia*, **26**, 1829–1841.
29. Shia,W.J., Okumura,A.J., Yan,M., Sarkeshik,A., Lo,M.C., Matsuura,S., Komeno,Y., Zhao,X., Nimer,S.D., Yates,J.R. III *et al.* (2012) PRMT1 interacts with AML1-ETO to promote its transcriptional activation and progenitor cell proliferative potential. *Blood*, **119**, 4953–4962.
30. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
31. Rustici,G., Kolesnikov,N., Brandizi,M., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Ison,J., Keays,M. *et al.* (2013) ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res.*, **41**, D987–D990.
32. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
33. Nottingham,W.T., Jarratt,A., Burgess,M., Speck,C.L., Cheng,J.F., Prabhakar,S., Rubin,E.M., Li,P.S., Sloane-Stanley,J., Kong,A.S.J. *et al.* (2007) Runx1-mediated hematopoietic stem cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. *Blood*, **110**, 4188–4197.
34. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Elsevier/Academic Press, Amsterdam; London.
35. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
36. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.